Big Data Over a 100G Network at Fermilab

Gabriele Garzoglio[1]**, Parag Mhashilkar, Hyunwoo Kim, Dave Dykstra, Marko Slyz**

Scientific Computing Division, Fermi National Accelerator Laboratory

E-mail: {garzoglio, parag, hyunwoo, dwd, mslyz}@fnal.gov

Abstract. As the need for Big Data in science becomes ever more relevant, networks around the world are upgrading their infrastructure to support high-speed interconnections. To support its mission, the high-energy physics community as a pioneer in Big Data has always been relying on the Fermi National Accelerator Laboratory to be at the forefront of storage and data movement. This need was reiterated in recent years with the data-taking rate of the major LHC experiments reaching tens of petabytes per year. At Fermilab, this resulted regularly in peaks of data movement on the WAN in and out of the laboratory of about 30 Gb/s and on the LAN between storage and computational farms of 160 Gb/s. To address these ever increasing needs, as of this year Fermilab is connected to the Energy Sciences Network (ESnet) through a 100 Gb/s link. To understand the optimal system- and application-level configuration to interface computational systems with the new high-speed interconnect, Fermilab has deployed a Network R&D facility connected to the ESnet 100G Testbed. For the past two years, the High Throughput Data Program has been using the Testbed to identify gaps in data movement middleware when transferring data at these high-speeds. The program has published evaluations of technologies typically used in High Energy Physics, such as GridFTP, XrootD, and Squid. This work presents the new R&D facility and the continuation of the evaluation program.

1. Introduction

Adaptation of 100GE Networking Infrastructure is the next step towards management of Big Data. The ability to efficiently store, retrieve, analyze and redistribute data is fundamental to scientific discoveries. Fermilab is the US Tier-1 center for the Large Hadron Collider's (LHC) Compact Muon Solenoid (CMS) experiment and the central data center for several other large-scale research collaborations. To deal with the scaling and wide-area distribution challenges of the data, the Laboratory is now connected to the next generation of networking infrastructure, ESnet's [1] 100G-backbone. This trend towards faster networks has driven the evolution of data management systems. Today, these systems provide access to and manage an increasingly distributed ensemble of resources, storing aggregate data approaching the exabyte scale.

To identify and address shortcomings in tools and technologies, Fermilab has devised a diverse program of work that spans all layers of computing to ensure full throughput in and across each layer. Figure 1 shows these research activities at different layers of computing. This program of work builds on core competencies on Big Data inherited from the RunII experiments at Fermilab and the CMS

---

[1] To whom any correspondence should be addressed.

expertise. It is a collaborative process that is part of the national program on Network R&D and involves multiple research organizations.

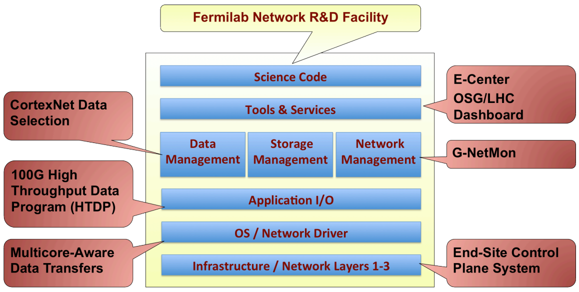


Figure 1. The Network R&D program at Fermilab spans all layers of computing to support the process of scientific discovery.

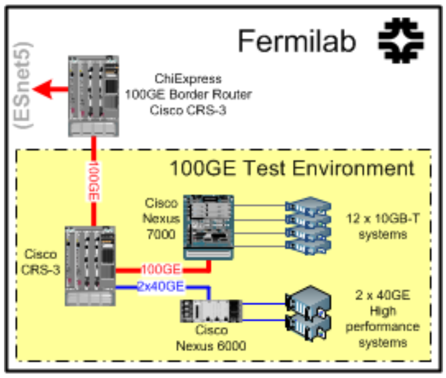


Figure 2. Fermilab Network R&D facility

To support this program, Fermilab has commissioned a dedicated R&D facility to do studies of 100G connectivity and technology evaluations in a production-like environment. Figure 2 shows a diagram of the facility.

At the core of the testbed there is a Cisco Nexus 7000 router supporting 100GE, 40GE, and 10GE ports. Attached to it, there are 12 nodes with 10GE network cards to achieve an aggregated throughput of 100 Gb/s. Another router, a Cisco Nexus 6000, is used to connect two nodes with 40GE cards for studies of high-performance systems, IPv6, load balancing, etc.

Part of the overall program of work is the High Throughput Data Program (HTDP), which focuses on ensuring that the data processing systems of the experiments are functional and effective at the 100G-scale end-to-end. We envision that HTDP will be a major user of the network R&D facility.

This paper is organized as follows. Section 2 and 3 highlight the work done in HTDP. We describe the problems identified while testing middleware on a 100 Gb/s network in section 4. Section 5 summarizes the results from our tests.

2. The High Throughput Data Program

The High Throughput Data Program (HTDP) at Fermilab focuses on the application I/O layer. HTDP aims at analyzing and tuning the performance of the end-to-end analysis systems used by High Energy Physics (HEP) experiments. To achieve this goal, the program performs and studies large data transfers over the 100GE ESnet networks. This work was fundamental in identifying gaps in the current data movement middleware services when used over 100GE networks.

HTDP is driven by the following three thrusts:

- Determine and tune the configuration of all layers to ensure full throughput in and across each layer/service.
- Measure and determine efficiency of the end-to-end solutions.
- Monitor, identify and mitigate error conditions.

In order to achieve these goals, we tested several data movement middleware services over high speed networking testbeds, such as the LIMAN testbed, Super Computing 2011 and ESnet’s 100 Gb/s testbed. Sections 2.1 and 2.2 below give an overview of these testbeds. The results from testing data middleware packages on them will then be presented in Section 3.

2.1. The LIMAN 30Gb/s Testbed & Super Computing 2011

The HTDP program started in 2011, participating in the Advanced Network Initiative (ANI) [2]. The ANI team deployed two fast-network test environments, in preparation for the full 100 Gb/s ESnet testbed. The first was the Long Island Metropolitan Area Network (LIMAN) testbed that connected BNL and New York at 30 Gb/s. The second environment was showcased at Super Computing 2011 [3], whereby a full 100 Gb/s network was made available for several communities to show the ability to utilize the fast network. The HTDP demonstrated the ability to transfer about 30 TB of CMS data in one hour using GridFTP [4] between NERSC and ANL with peak rates of 75 Gb/s. Details on the testbed configuration, test cases, and the test results can be found in [5][6].

2.2. The ESnet 100 Gb/s Testbed

In 2012, ESnet deployed a 100 Gb/s network testbed connecting NERSC and ANL. Each of these sites hosted three high performance servers connected via two 100 GE border routers at each site. The description of the testbed and the machines can be found in [5].

3. Data Middleware Validation on the ESnet 100G Testbed

Following the methodologies developed for the LIMAN testbed, we conducted similar tests on ESnet's 100 Gb/s testbed for GridFTP, Globus Online (GO) [7], Squid [8] and XrootD [9] to saturate the 100 Gb/s network with different file sizes. This section describes the tests conducted and their results.

3.1. GridFTP & Globus Online

In comparison to simple bandwidth measurement utilities such as nuttcp, GridFTP introduces additional overheads due to the increased security and reliability. High-bandwidth utilization for medium (4 MB to 1 GB) and large (2 GB to 8 GB) files was achieved by tuning the GridFTP transfer parameters. In particular as shown in figure 3, for our round trip time (RTT) of 53 ms, we achieved bandwidths above 90 Gb/s with a concurrency and a parallelism set to 4 (-cc 4 -p 4), using the default TCP window size. Tuning the same parameters, however, was not sufficient to achieve similar high utilization when transferring small files. As described in Section 4, we had to introduce GridFTP pipelining to improve the performance and address the *Lots Of Small Files (LOSF)* issue.

Further investigations revealed that logging exposed an additional impact in the measured bandwidth for small files. Since typically every transfer event is logged, this resulted in per-file access to a system disk. This was especially harmful to performance because the ESnet testbed configuration mounted the system disk remotely over a 1 Gb/s network. The blue and red bars in the small file sizes section of figure 3 shows the effect when logging was turned on (blue bar) or off (red bar). The lesson learned was the need to understand every single detail of a testbed to take credible measures.

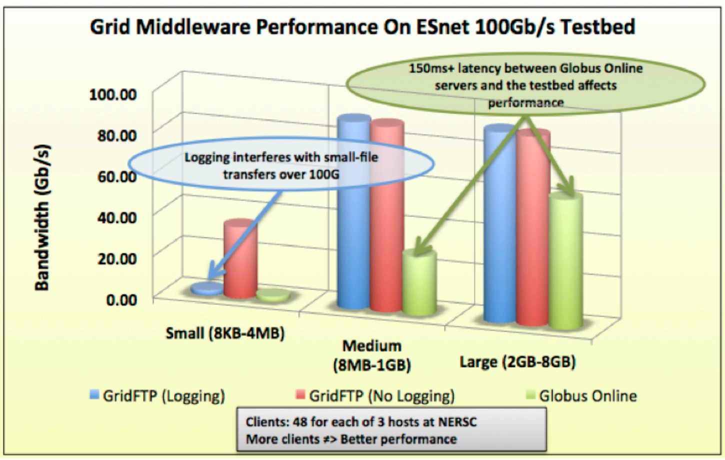


Figure 3. Bandwidth measurements for GridFTP and Globus Online over the ESnet 100 Gb/s testbed

We did similar measurements using GO with manual and auto-tuning of the transfer parameters. Because of high latencies in forwarding the control channel from GO to the ESnet testbed VPN through a Fermilab node, GO showed a poorer performance, compared to GridFTP, at ~30 Gb/s for medium file sizes and 60 Gb/s for large sizes.

3.2. XrootD & Squid

Along with testing GridFTP and Globus Online, we also tested the performance of other commonly used Grid middleware like XrootD and Squid. The test setup and results are described in [6] in detail. This section shows the performance observed and summarizes the results.

Figure 4 shows the performance of XrootD on the ESnet testbed. The XrootD server performed better with increasing numbers of clients, achieving almost 90% bandwidth utilization for 8 clients.

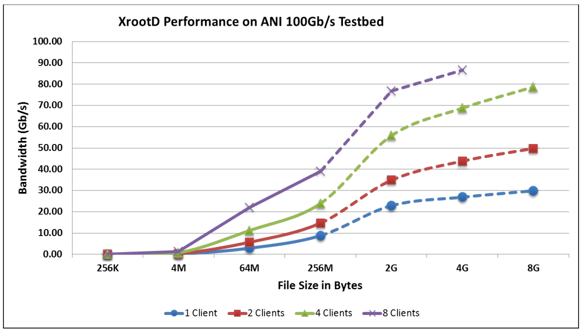


Figure 4. XrootD performance on ESnet 100G Testbed

Figure 5 shows the performance of Squid on the ESnet 100G testbed with and without core affinity (CA) for different numbers of clients per machine (CPM). We tested the following client/server configurations:

- One-to-one transfer: One client on a host at NERSC requesting 8MB file from exactly one server on a host at ANL and vice versa.
- All-to-all transfer: Each of the clients at NERSC requesting 8MB file to all the servers at ANL and vice versa.

For the tests from ANL to NERSC, one-to-one showed lower performance than all-to-all. One possible reason was that one of the NERSC machines was slower than the others and had one 10 Gb/s NIC with frequent problems. This decreased the utilization with more client load. For the tests from NERSC to ANL, one-to-one performance was almost the same as all-to-all. When the servers were running at NERSC, even the one bad NIC did not cause any issue because the other three NICs were compensating by sending more data.

For both direction benchmarks, it is worthwhile noting that the core-affinity optimization was able to improve the overall performance by up to 21%. Furthermore, increasing the number of Squid servers per machine also improved aggregate performance, with some speeds approaching 100 Gb/s.

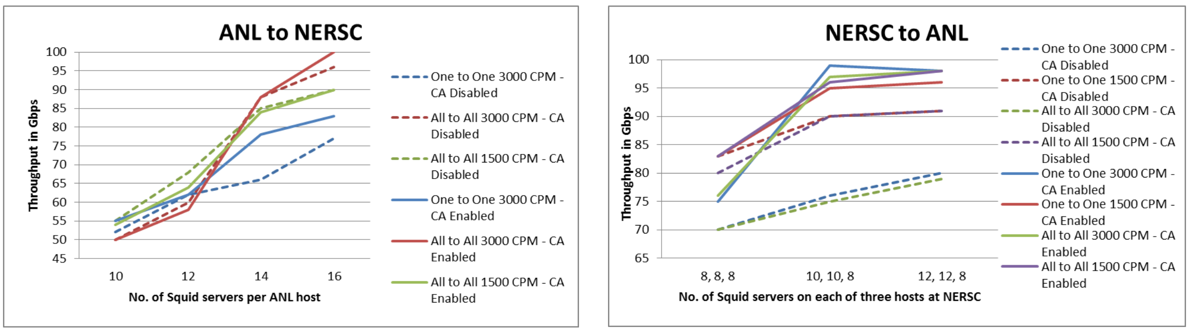
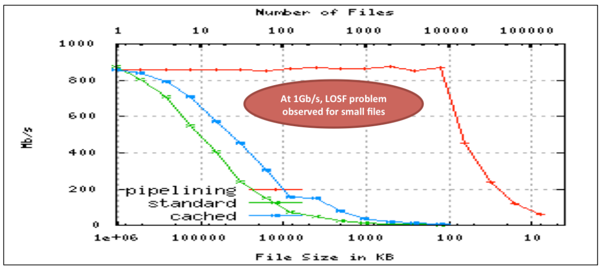


Figure 5. Squid performance on ESnet 100G Testbed

4. LOSF & Middleware Performance

The experimental results from testing data transfer applications, such as GridFTP and Globus Online, revealed that over 90% bandwidth utilization can be achieved on average when transferring datasets with medium (8MB – 1GB) and large (2GB – 8GB) file sizes. When we did the same study at a finer granularity of file sizes, however, this higher bandwidth utilization was consistently achieved only when transferring files 32MB or larger. File sizes smaller than 32MB are affected by the Lots of Small Files (LOSF) problem, which typically results in significantly lower transfer bandwidth utilization. This is due to the OS, system and protocol-level overheads present when reading and transferring individual files.

Figure 6 shows that the LOSF threshold depends on the network bandwidth. Pipelining in GridFTP contributes to reducing the LOSF threshold because subsequent files in a pipeline are transferred without waiting for the successful transmission of the previous file, thus reducing one round trip time (RTT) delay between two consecutive files transferred.



Reference: John Bresnahan, Michael Link, Rajkumar Kettimuthu, Dan Fraser, Ian Foster; "GridFTP Pipelining", In the proceedings of Teragrid 2007 Conference, Madison, WI, USA

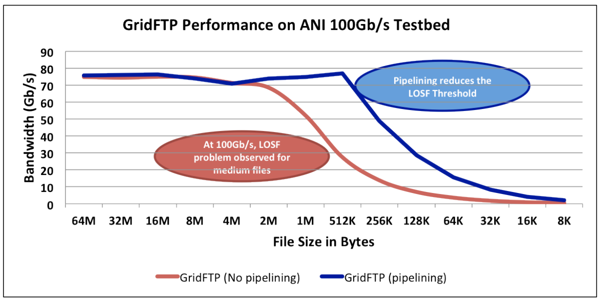


Figure 6. Lots Of Small Files (LOSF) Problem on 1Gb/s v/s 100Gb/s

5. Summary

The Network R&D program at Fermilab is driven by the needs of the High Energy Physics and Astrophysics stakeholders. The program spans all layers of the communication stack. To provide an infrastructure to conduct the R&D program, Fermilab has deployed a Network R&D testbed connected to the 100 Gb/s ESnet backbone.

At the application layer, the High Throughput Data Program (HTDP) studies the end-to-end functionality and performance of major analysis systems of scientific stakeholders on 100 Gb/s networks. Recent work includes studies of data transfers with applications such as GridFTP and Globus Online on the ESnet 100G testbed between NERSC and ANL.

Both GridFTP and Globus Online showed good performance over large and medium size files. The Lot Of Small Files (LOSF) problem impacts the performance transferring small size files. XrootD servers showed better network utilization with the increase in number of clients. In the case of Squid, core-affinity considerably increased the bandwidth utilization.

The network R&D program will continue in 2014 with the validation of the new 100 Gb/s capabilities at Fermilab.

Acknowledgements

Fermilab is operated by Fermi Research Alliance, LLC under Contract number DE-AC02-07CH11359 with the United States Department of Energy.

References

[1] Energy Sciences Network. Accessed on Oct 16, 2013. http://www.es.net
[2] The Advanced Networking Initiative. Accessed on Oct 16, 2013. http://www.es.net/RandD.
[3] SuperComputing 2011. Accessed on Jun 1, 2012. http://sc11.supercomputing.org/
[4] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, I. Foster, The Globus Striped GridFTP Framework and Server, in Proc. of the 2005 ACM/IEEE conference on

Supercomputing, pp.54-64, Seattle, Washington USA, November 2005.

[5] Dykstra, Dave, Gabriele Garzoglio, Hyunwoo Kim, and Parag Mhashilkar. "Identifying Gaps in Grid Middleware on Fast Networks with the Advanced Networking Initiative." In Journal of Physics: Conference Series, vol. 396, no. 3, p. 032034. IOP Publishing, 2012.

[6] Rajendran, Anupam, Parag Mhashilkar, Hyunwoo Kim, Dave Dykstra, Gabriele Garzoglio, and Ioan Raicu. "Optimizing Large Data Transfers over 100Gbps Wide Area Networks."

[7] Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing (May/June):70-73, 2011.

[8] D. Wessels, Squid: The Definitive Guide, O'Reilly & Associates, Inc. Sebastopol, CA, USA ©2004, ISBN:0596001622

[9] A. Dorigo, P. Elmer, F. Furano, and A. Hanushevsky. XROOTD-A Highly scalable architecture for data access. WSEAS Transactions on Computers, 1(4.3), 2005.